Collated Notes on ET-NCMP software

Simon Grainger

Thank you for testing the code and your thoughtful comments.

I agree that the problem of missing data will be a major challenge. I have looked some more at the computation of the monthly indices. I think for each station it is possible to compute monthly indices continuously from January of the first year to December of the last year of data, with missing months (and years) filled in by missing values. However, the consequences of this approach will have to be carefully tested.

Documenting the warnings and error messages that you found will be very useful in trying to improve the code

Thank you for your comments about the R-NCMP software. From them, and my own testing, I think that there are two main areas in the software which should be addressed prior to releasing the R-NCMP to a broader user group. They are:

1. What recommend/default values should be specified for a range of analysis parameters, and how difficult should it be for a casual user of the code to change those parameters. (Since it is Open Source, it is not possible to prevent expert users modifying R code.)

2. What is the best approach for treating missing data particularly, but not limited to, in the calculation of extremes indices.

I plan to take a look at the second point this week. I know that there are issues in how the monthly diagnostics are calculated from the daily data, but I am not completely sure how the code responds to large amounts of missing daily values. The first point will require discussion by the Expert Team.

Ladislaus Chang’a

Thank you for the feedback. I have also tested the software for Tanzania Data and the following are my observations and comments

1. The Software run very fast

2. We tried to include as many stations as possible that meant including stations with many missing data and the results was several Warning Messages and in some occasion the Program stopped particularly when computing the Indices

3. We Later tried to remove stations with a lot of missing data and the program run smoothly with few Warning Messages and few error message

4. Computation of the Variogram using All options could not go well and instead we computed the Variogram by running for one NCMP at a time which worked

5. So we Think Treatment of the Missing Data is the one of the major challenges that need to be addressed and get more clarity including understanding the criteria for the Number of missing data for the station to be excluded from the analysis

6. It will be useful to document all the Potential Warning and Error messages and the guidance on how to resolve it

These are few observations on our side

Lucie Vincent

Comments and questions on the R-NCMPs software – January 31, 2018

P1\_Quality\_Control.R

1. I am wondering if it would be preferable to have a user defined upper and lower limits to identify outliers in tmax and tmin. Currently, outliers are identified when tmin < -65°C and tmax > 50°C.
2. Outliers are also identified when tmax (or tmin) is outside of mean + nSD. ET recommends of using 5 SD: however 5 SD identified a lot of values. Should ET recommend using 7 SD instead?

P2\_Indices.R

1. The criteria for calculating monthly and annual values have changed:
2. monthly values are calculated if # of missing days ≤ 10 days (WMO 2009)
3. annual values are calculated if # of missing days ≤ 36 days (WMO 2009)
4. climatology is calculated if there is at least 9 valid years??? (WMO 2009)

(We need the complete reference for WMO 2009)

2. There is a new calculation done for NCMP6: how is it done? The record is calculated from the first 30 years of data: what if the station length less than 30 years?

P3\_Variogram.R

1. What is the minimum number of years recommended by ET to produce the variogram? Do we want to suggest using 1 year only? Which year should be used? Are variograms better when more years are used? ET should provide some guidance in the User Manual.
2. Is the procedure described in Appendix D still valid? Are the points still divided into bins of 20 km and the maximum distance is 3000 km for temperature and 2000 km for precipitation? The interactive program to produce the variogram is helpful!
3. Does this procedure work for small countries and small islands? Maybe more testing needs to be done.

P4\_Region\_Average.R

1. After entering the 3-digit code, it is mentioned “Calculating variograms for all stations with computed indices. Can change this by editing the file ‘A2\_Indices/P2\_Station\_List.txt”. What does the message mean? Are the variograms re- computed in this program? Why?

P7\_Summary.R

1. A file named “CountryName\_NCMP\_Summary.csv” should be created in the folder A7\_Summary. The user would have to enter the name of the country or region (ex. Canada) and the year for producing the summary (ex. 2015). The columns would include:
2. Country
3. Year
4. Month (or annual value)
5. Country average of TMA, rank, number of station contributing
6. Country average of PrAn, rank, number of station contributing
7. Country average of SPI, rank, number of station contributing
8. Country average of TX90p, rank, number of station contributing
9. Country average of TN10p, rank, number of station contributing
10. Base period start (ex. 1981), and base period end (ex. 2010).

The output would be for 13 rows (12 months and annual) and 20 columns.

Fatima Driouech

I have done calculation with Moroccan data (30 stations) using the new code. It is quite fast and relatively smooth .

In addition to several warnings, I noted the folowing:

- generation of "Inf" value for Index 2 in stations with very few precipitation (i.e in summer) and this not allow progression in calculation (variorgam, ...),

I solved the problem by putting -99.9, but this is not a durable neither good solution.

- When asking calculations for only few stations (among the list included in P0\_Station\_List or P2\_Station\_List), we can not select the stations. Calculations are done starting from first station in the list.

- When using the P3\_Variogram\_Interactive program, I didn't find the intermediate variograms to check if they are correct or no; may be because they are printed into the output file at the end (or I missed the right folder).

I hope this points can be solved easily.

James Adams

The requirement of a certain version of R required a downgrade or parallel installation of R, in that I already had another, more recent version installed. My recommendation is to make sure that the code works for any version of R after a base level and state this requirement, i.e. “this code should run on R version 3.3.1 or greater”. This will lighten the burden for users who already have R installed.

When installing the required packages I received the following warnings:

Warning messages:

1: package ‘stats’ is not available (for R version 3.3.1)

2: package ‘stats’ is a base package, and should not be updated

I was never prompted for a personal library download location as described in the last of part of section 2.2 of the user manual.

In section 2.4 of the user manual it mentions that you need to make sure you don’t have anything in the directory that will be overwritten. This section doesn’t allude to what those files are, I assume they’re described later in the document. My take is that the code should be modified so this is not a potential issue, as users will invariably make this mistake at some point. (In general if you foresee users making an error but you can code around it to shield them then it’s advisable to do so.) The names of the files to be concerned about should be itemized in section 2.4 of the user manual.

The requirement of a 27 character file name for the inputs seems arbitrary/inflexible/unnecessary. Is there a good reason for this? Again if flexibility can be cooked into things then users are less likely to shoot themselves in the foot.

Perhaps make default values for user prompts so the user can just hit Enter to select the default value, especially in cases where there is a recommended option, such as 1981 for the initial year of the base period.

Everything was smooth sailing until reaching the variograms calculation stage. There were numerous warnings:

Variograms Done!

There were 50 or more warnings (use warnings() to see the first 50)

> warnings()

Warning messages:

1: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

2: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

3: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

4: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

5: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

6: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

7: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

8: In nls(Bl ~ Gaussian(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

9: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

10: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

11: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

12: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

13: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

14: In nls(Bl ~ Gaussian(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

15: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

16: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

17: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

18: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

19: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

20: In nls(Bl ~ Gaussian(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

21: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

22: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

23: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

24: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

25: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

26: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

27: In nls(Bl ~ Gaussian(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

28: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

29: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

30: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

31: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

32: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

33: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

34: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

35: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

36: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

37: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

38: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

39: In nls(Bl ~ Gaussian(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

40: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

41: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

42: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

43: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

44: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

45: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

46: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

47: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

48: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

49: In nls(Bl ~ Spherical(Dl, a, b, c), start = list(a = 0.01, ... :

 singular gradient

50: In nls(Bl ~ Exponential(Dl, a, b, c), start = list(a = 0.01, ... :

 step factor 0.000488281 reduced below 'minFactor' of 0.000976562

Region averaging went fine with only a couple of warning messages:

Warning messages:

1: package ‘maptools’ was built under R version 3.3.3

2: package ‘sp’ was built under R version 3.3.3

The trends and graphs went fine as well, with similar warnings:

Warning messages:

1: package ‘maps’ was built under R version 3.3.3

2: package ‘zyp’ was built under R version 3.3.3

3: package ‘Kendall’ was built under R version 3.3.3

John Kennedy

Overall the software worked really nicely. Thanks to Megan, Lucie, Simon and Justin for all their work on this.

General

These are mostly cosmetic or user-experience type comments.

It might be nice to make the questions and messages in the software less terse (or more verbose depending on your point of view). For example, when a choice is offered to the user, we indicate the recommended value by saying “ex. 300”. We could say “(recommended 300)”.

We could arrange it so that when the script starts it tells the user what it does with a very short description. e.g. for P2\_Indices.R it could say:

“This script calculates the ETCCDI indices that are needed for the National Climate Monitoring Products for each of the input stations.”

P3\_variograms already does this, which is nice.

Specifics

If we could arrange the default value to be chosen if there is no input (i.e. the user just hits return), then that might make it even easier for people to choose the “best” values.

Missing data criteria were a bit strict for the annual values for the UK. I changed the maximum number of missing days in a year to be 120 and that fixed it. This might worry some people. Making the number of days an input would be helpful as we are not prescriptive in the guidance about this.

When I originally ran the code, all the annual values were missing for some station and this caused the code to fail. Some basic data checks in the scripts would be useful to make sure that the data are digestible.

Outputs. The code already produces a set of output plots which are really useful. At the moment, the only output I see is annual and monthly plots may be of use for tracking down odd things and looking at seasonal variations.

Outputs: monthly and annual maps of the interpolated index would be a useful output although we might have to make them small so that they don’t use up huge amounts of disk space. At the moment, running for the UK at 0.1° resolution gives a good estimate of the regional average, but uses about 1.2Gbytes of disk space, mostly for the grids. This isn’t a huge amount by modern standards, but it’s enough to cause problems.

For the variogram calculation, we should set the recommended values to 1981 and 2010 for the start and end of the period used to estimate the variograms.

Small countries

We need to test this. In the simple case where we have one station, then the calculation should default to using only that station.

Where there are a small number of stations, then there ought to be a default variogram because calculating a variogram would not be very robust. This would necessitate adding an extra check and question. If there are fewer than 10 stations, then you have the option (recommended) of using a default variogram.

An alternative is to allow stations from outside the country to be used to improve the estimate. As I understand it, at the moment, the code will accept stations that are outside the country limits and these can help to better define the interpolated indices at the borders. I filtered GHCN-Daily into the file format required by the NCMP code and I’ve used this to test the NCMP in non-UK areas, so we could potentially use something like this to supplement in-country data. We would need to exclude GHCND data from within the country somehow.